CWI Quarterly Volume 12 (2) 1999, pp. 147 – 157

Cross-Language Multi-Media Information Retrieval

Páraic Sheridan, Peter Schäuble
Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich, Switzerland

We describe here the evaluation of a cross-language information retrieval technique based on *similarity thesauri* in a multi-media document environment, such as is likely to be found in a digital library. We present the theory of similarity thesauri, which are information structures derived from corpora, and show how they can be used for cross-language retrieval. In evaluating our similarity thesaurus-based approach to cross-language retrieval over a parallel collection of legal texts, we show that cross-language retrieval can perform *equally as well as* monolingual retrieval in the certain cases. We also present the results of a first evaluation of cross-language retrieval with spoken news material. We conclude that providing cross-language access to multi-media digital libraries is already a viable possibility.

1. Introduction

There are many example scenarios in which users of a digital library may be interested in information which is in a language other than their own native language. We believe that in many cases, a common language scenario is where a user has some comprehension ability for a given language but is not sufficiently proficient to confidently specify a search request in that language. In this case, a search system which accepts search requests in one language (the user's preferred language) and retrieves relevant information in other languages (the languages of the digital repository) is of benefit. This is the process of *cross-language information retrieval*.

As digital information repositories grow ever larger, they will increasingly move beyond the textual paradigm and include more and more speech and video material, creating new challenges for information retrieval [7]. Speech and video information is also language-specific and must also therefore be taken into account when considering issues of multilingual and cross-language access.

In this paper we consider both textual and spoken information when examining cross-language information retrieval. Our approach is based upon the use of information structures known as *similarity thesauri* which are corpus-based and can be constructed over any multilingual collection of comparable documents. The theory underlying similarity thesauri is presented in section 2. We have used a collection of Swiss legal texts to construct an evaluation environment in which we have tested the effectiveness of our cross-language retrieval system, the results of which are presented in section 3. Section 4 then presents our work on cross-language speech retrieval, including an evaluation over a collection of German radio news. We conclude in section 5 with some comments on our results and on the outlook for the future.

2. Similarity Thesauri

A similarity thesaurus is an information structure representing term similarities which reflect domain knowledge of the collection over which the thesaurus is constructed. The term similarities recorded within a similarity thesaurus are determined based on *how the terms of the collection are indexed by the documents*. The theory underlying the construction of similarity thesauri is therefore probably best understood by thinking, in information retrieval terms, of exchanging the roles of documents and terms in the traditional view of document retrieval. The documents serve as indexing features and the terms represent retrievable items.

The idea of similarity thesauri was first developed with the aim of facilitating query expansion for monolingual retrieval [10]. A similarity thesaurus was constructed over the retrieval collection so as to capture certain domain knowledge through the term similarities within the collection. Expanding user queries by extracting from the similarity thesaurus the most similar terms to the query concept lead to improvements of up to 51% in retrieval performance (average precision non-interpolated) on a 2.3 GByte collection of TREC documents [9]. We have recently expanded the use of similarity thesauri, noting that in applying similarity thesaurus technology to collections of multilingual documents we could identify cross-language term similarities, thereby capturing a *translation effect* within the similarity thesaurus. Instead of straight query expansion, we can then use a multilingual similarity thesaurus for pseudo-translation of user queries, thereby facilitating cross-language retrieval [12]. A much more detailed and formal presentation of the theory underlying similarity thesauri can be found in [14].

3. Cross-Language Text Retrieval

In our current evaluation of cross-language text retrieval we have been working with two documents collections from the Swiss legal domain. The first is a parallel collection in French, German and Italian of the Swiss federal law, totalling about 155,000 documents. The second collection is made up of the decisions of the Swiss federal court of justice since 1975, in which documents

are in either French, German or Italian, totalling about 8,000 documents.

What is particularly useful for our experimental purposes is the fact that the collection of decisions of the federal court includes a German keyword index with references from the index to the relevant articles of Swiss federal law. This allows us to use both collections in combination as a test environment for retrieval performance, with the keyword index of the federal court decisions as the query set and the linked articles of Swiss law (and all sub-articles) as the relevant document sets.

The evaluation of information retrieval systems has traditionally been accomplished through a pair of measures based on the relevance of retrieved documents to the query: *recall*, measuring the portion of relevant documents that were retrieved by the system, and *precision*, measuring the portion of retrieved documents that are actually relevant. The relevance of documents to queries is usually determined by domain experts. This task of judging relevance of documents to queries however, requires enormous resources, especially given the number of documents and queries that are considered necessary for evaluation of modern retrieval systems. In order to overcome this resource requirement, especially when trying to evaluate cross-language retrieval systems, researchers have had to either adopt different evaluation measures [5] or find novel ways of approaching the relevance judgement task [13]. In the Swiss legal collections however, the legal experts who constructed the keyword index linking the decisions of the federal court to the relevant passages of Swiss law have, in effect, already performed the relevance assessment task needed for an information retrieval evaluation.

There are 105 entries in the German keyword index of the decisions of the federal court of justice which have links to articles of Swiss law. Most entries (84) are single-word terms, though there are also some two- and three-word terms. These 105 entries therefore correspond to 105 test queries used in evaluating our retrieval system. The majority of keyword entries point to only one article of the Swiss law, though some also reference two or three articles.

Because of the hierarchical nature of the legal texts reflecting the division of the law into articles and sub-articles and so on, a reference to a given legal text from the decisions of the federal court did not reflect the relevance of only a single document to that keyword (the document referenced in the index), but also all documents subordinate to the referenced document in the hierarchy of the law. For example, if the keyword index of the federal court decisions refers to document *SR 443.1* of the Swiss law, then all documents subordinate to this in the law (e.g. *SR 443.1.10, SR 443.1.5.2, SR 443.1...*) are considered relevant to that keyword as a query. Using this method of assessing relevance, the 105 test queries have on average 252 relevant documents each.

The purpose of our evaluation here is to determine the effectiveness of our similarity thesaurus approach to cross-language information retrieval compared to a baseline performance set by an equivalent monolingual retrieval task. The core of our evaluation is therefore the comparison of the performance of German monolingual retrieval against the performance of German queries retrieving

French documents. Since the document collection of the Swiss federal law is parallel, the precision/recall figures for the monolingual and cross-language retrieval tasks can be directly compared.

3.1. Experimental Results

The baseline monolingual retrieval task, against which cross-language retrieval performance was to be compared, was simply the retrieval of German texts from the Swiss federal law collection given 105 German queries from the keyword index of the Swiss federal court of justice. Average precision over the 105 queries was 0.2863. To compare cross-language retrieval against this baseline we used a similarity thesaurus to retrieve French texts from the Swiss federal law collection for the 105 German queries of the federal court index. The similarity thesaurus was constructed using the French/German parallel collection of the Swiss law. This resulted in a 64 MByte similarity thesaurus containing 91,310 German terms and 43,066 French terms. Choosing the 5 most similar French terms from this similarity thesaurus for each German query was empirically determined to give the best cross-language retrieval results, with an average precision of 0.3297. Surprisingly, this represents a *15% improvement* over the performance of the monolingual baseline.

Given this result, we felt that explanation for the better performance of cross-language retrieval compared to the monolingual case was most likely to be found in the expansion effect of the similarity thesaurus used in creating a pseudo-translation of the queries. Remember the majority of queries (84) are only one word. The expansion effect would be in keeping with the substantial performance improvements found when similarity thesauri were originally used in for monolingual query expansion. This hypothesis was easily tested; by constructing a monolingual German similarity thesaurus over the Swiss federal law collection and rerunning the baseline experiment using the German similarity thesaurus to expand the German queries.

This experiment confirmed our hypothesis about the benefit of query expansion. Expanding the German queries to five terms results in a 17% improvement in average precision (0.3355) over the original monolingual baseline and leaves a difference of less than 2% between the performance of monolingual (expanded) and cross-language retrieval.

These experiments serve to demonstrate the usefulness of similarity thesauri, both reinforcing their use for query expansion in monolingual retrieval, and more importantly as an approach to cross-language information retrieval. It is important to note however that the evaluation scenario presented here so far has been ideal, in that the similarity thesaurus has been constructed over a parallel document collection, the same collection as was used in the retrieval experiments. It is more likely in practice that cross-language retrieval be based on the use of a similarity thesaurus that has been constructed over some separate *training* collection.

To create such a more realistic evaluation, we constructed a similarity the-

saurus over the parallel one-paragraph summaries of the decisions of the federal court of justice. The similarity thesaurus constructed over this collection is significantly smaller than the one from the corpus of Swiss federal law, since there are only 8,282 documents in the corpus and each document is only one paragraph. On the other hand, there is still a very close domain match between decisions of the Swiss federal court of justice and the texts of the Swiss federal law. The results of re-running our cross-language retrieval experiments, retrieving French documents of Swiss law from German queries, are presented in Figure 1, which also contains our previous results for comparison.

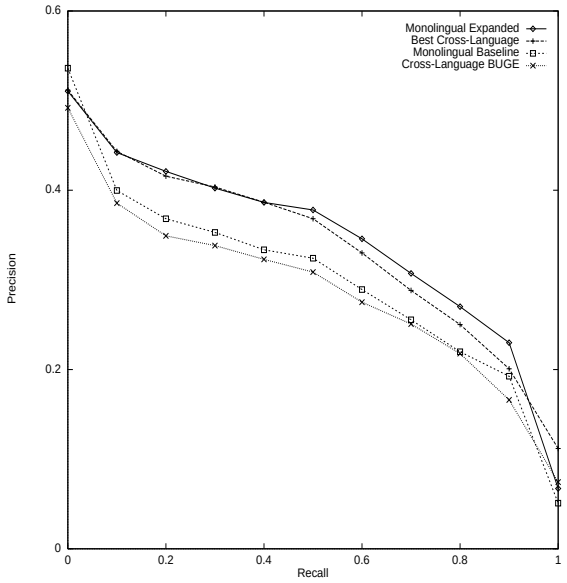

FIGURE 1. Overview of experiment results

As expected, the use of a similarity thesaurus trained over a document collection different than the one used for retrieval leads to a certain degradation in performance. Compared to the performance of cross-language retrieval using the similarity thesaurus of the Swiss federal law collection, using a similarity thesaurus from the decisions of the federal court of justice (denoted as 'BUGE' in our figures here) results in a 16% loss in performance. This equates to a level of performance 18% below that achieved on the monolingual task when a monolingual thesaurus was used for German query expansion. A summary of average precision values for each of our experimental tests are given together in Table 1

	Avg. Prec.	Diff.	
Monolingual Baseline	0.2863		
Monolingual Expanded	0.3355	+17%	
Cross-language (5 terms)	0.3297	+15%	(-2%)
Cross-language (BUGE)	0.2749	-4%	(-18%)

TABLE 1. Comparison of Experimental Results

4. CROSS-LANGUAGE SPEECH RETRIEVAL

Having addressed independently in the past the problems involved with content-based retrieval of speech data [16], we recently turned our attention to combining our research in cross-language retrieval and speech indexing and retrieval to investigate the viability of cross-language speech retrieval [15]. Our aim was to establish a *baseline* of performance on this task, against which we can then measure the success of our continuing research in this area.

In our approach to speech retrieval, indexing is based on phonemic transcriptions determined by a phoneme recogniser. The phonemic transcriptions are indexed using overlapping N-gram features. At retrieval time, an additional probabilistic matching technique may be applied, during which individual words are matched *fuzzily* against the erroneous transcriptions. This technique has proven to be effective not only in speech retrieval [17] but also in retrieval from error-prone OCR texts [6]. This approach to speech retrieval has the substantial advantages of requiring only a relatively simple phoneme recogniser and having a theoretically unrestricted search vocabulary (the words of a language are composed from a closed set of phonemes). The only practical restriction on search vocabulary comes from the dictionary which is required to translate query words into their phonemic transcriptions.

The speech retrieval module used here is based on a *speaker-independent* phoneme recogniser for German speech which we have constructed using the HTK toolkit [18] and trained on 3:44 hours of the PhonDat speech corpus [8]. In an evaluation of this phoneme recogniser on a very small test set based on radio news (9 speakers, 5 minutes), only 49% of the phonemes were detected correctly, whereas state-of-the-art phone or phoneme recognisers can operate at recognition rates up to 75%. The poor performance of our recogniser may be attributed to the fact that the PhonDat training set is so different from our audio news test collection. For example there are no speakers common to

the two collections and the speakers' dialects in the PhonDat set correspond to regions of Germany, not Switzerland. Techniques like speaker adaptation or model refinement by triphones could be applied to improve our recognition quality.

The input to the speech retrieval module takes the form of German text queries. In a first step, the query features (non-stopword stems) are phonemically transcribed using a phoneme dictionary. The phoneme dictionary used here consists of 373,000 entries and consists mostly of the dictionary of the CELEX-2 CD-ROM [4]. To facilitate cross-language speech retrieval, we used a French/German similarity thesaurus which was constructed over a corpus of French and German news stories from the Swiss news agency (SDA) for the years 1988, 1989 and 1990, totalling about 83,000 documents.

For evaluating the effectiveness of speech retrieval, our speech collection consisted of approximately 30 hours of Swiss radio news covering the time range from April to December 1995. This news material was collected automatically using a system set to record 7 minutes per day beginning at 7am or 9am. There are at least 10 different speakers, both male and female, represented in the collection. Another interesting feature is the fact that news bulletins often include reports from correspondents over the telephone line. In general, a recorded news bulletin covers approximately 5 to 7 different news stories, and story boundaries can not be automatically determined. For retrieval, we cut all recordings into non-overlapping fixed length documents of 20 seconds duration, regardless of story boundaries. This resulted in a collection of 5,397 audio documents.

The queries for our evaluation were collected from independent sources; year-end news reviews published in several different Swiss newspapers at the end of 1995. These reviews contained a brief summary of each major news event during that year. We extracted 26 queries based on the summaries of events dated in the range October to December 1995. The average query length is 10.5 terms. For the cross-language experiments, these German queries were manually translated into French.

4.1. Experimental Results

The main objective of our experiments was to compare the performance of a cross-language speech retrieval task with the baseline monolingual speech retrieval. The monolingual baseline was established by submitting the original German queries derived from the newspaper reviews to the speech retrieval module. This experiment was labelled *DE_Base*. The basic cross-language experiment involved submitting the manually translated French queries to the similarity thesaurus to produce German pseudo-translations, and these pseudo-translated German queries were then submitted to the speech retrieval module. This experiment was labelled *FR_Sim_Base*. Query pseudo-translation was performed by translating each input French query term with the two most similar German terms stored in the similarity thesaurus.

Two further experiments were performed to test the effectiveness of query expansion based on an automatic relevance feedback loop over an independent but similar document collection, a refinement that has been found to be useful for cross-language retrieval [1]. The first of these tests was aimed at establishing the usefulness of this relevance feedback loop as a query refinement step performed after query pseudo-translation. The input French queries were first pseudo-translated as in the baseline cross-language experiments. The resulting German pseudo-translations were then submitted as queries to a collection of Swiss newspaper (NZZ) stories from 1995. The top 5 documents retrieved were automatically assumed to be relevant and query term re-weighting was then performed using the Rocchio formula [11]. The top weighted 10 features were then returned as the output query to be submitted to the speech retrieval module. This cross-language run with feedback is labelled *FR_Sim_NZZ*.

We tested similarly the usefulness of the automatic feedback loop as a query expansion step prior to speech retrieval in the monolingual case. The original German queries were evaluated against the newspaper text collection as a first step and automatic relevance feedback performed as described above. Again, the top 10 re-weighted features were then submitted to the speech retrieval module. This is labelled as *DE_NZZ*. The results of these four experiments are presented in Figure 2.

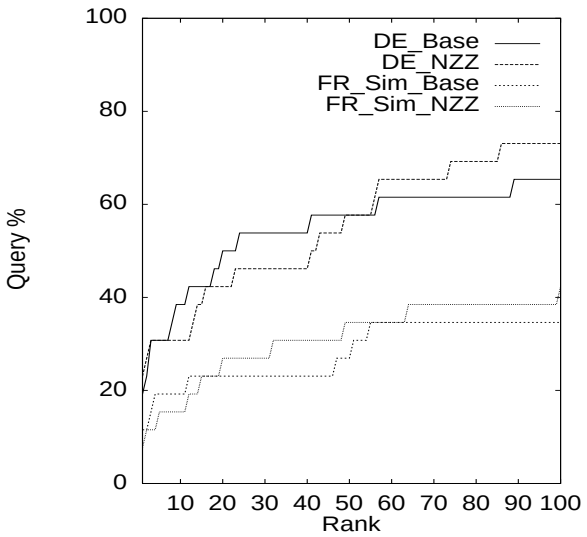

FIGURE 2. Results using tri-gram matching only

Our evaluation strategy is precision oriented. For each query, we measure the rank position of the highest-ranked relevant document. This strategy was selected in order to avoid the huge effort needed in gathering the relevance

information required for a standard IR evaluation based on precision and recall. The result graphs presented here therefore show the percentage of queries where a relevant document was found within a given range of the ranked list of documents retrieved. In our experiments we considered only the top 100 retrieved documents for each query. Figure 2, for example, shows that for 57% of queries, a relevant document was found within the first 50 ranks for the monolingual baseline experiment *DE_Base*, whereas for the cross-language baseline *FR_Sim_Base*, only 27% of the queries had a relevant document returned within the first 50 retrieved documents.

To compare the performance of different variations in cross-language and monolingual speech retrieval using the results presented here, we can choose a rank position and examine the percentage of queries which returned a relevant document above that position. Since we are interested in high-precision performance we can compare performance at the top 10 document cut-off level. Given this measure, the baseline monolingual speech retrieval run performs best retrieving a relevant document in the top 10 for almost 40% of the queries. The use of an initial feedback loop for query expansion has not helped in this configuration, returning a relevant document in the top 10 for just over 30% of the queries. In the cross-language experiments, the baseline is at 20% of queries and the query refinement loop using automatic feedback is again worse, with a relevant document in the top 10 for around 18% of queries.

Although there is an obvious need for further evaluation of our approach to cross-language speech retrieval, these initial experiments allow us to comment on a baseline performance for this task, as follows:

20% of queries have at least one relevant document in the top 10 documents returned by cross-language speech retrieval, compared to 44% of queries in the monolingual case, when evaluated over 26 French queries submitted against 5,397 20-second German audio documents derived from 30 hours of radio news.

On the face of it, this level of performance is quite poor. For every second monolingual query there are no relevant documents in the top ten, and only one in five cross-language queries retrieves a relevant document in the top ten, though this is based on a relatively small query sample. What we are interested in here however, is the performance of *cross-language* speech retrieval *relative* to monolingual speech retrieval. Cross-language speech retrieval is performing here at about 45% of the monolingual speech retrieval performance. Interestingly, this is not far off the 50% performance relative to monolingual retrieval reported with some basic approaches to cross-language *text* retrieval [2], [3] (though these authors have also achieved better performance than this). We are satisfied that this is an acceptable starting baseline for performance of cross-language speech retrieval.

5. Conclusion

The evaluation we have presented here firmly establishes the usefulness of similarity thesauri for providing cross-language access to multi-media digital repositories. Similarity thesauri bring the additional advantages of query expansion

to the cross-language task, the extent of which is clearly demonstrated in our experiments in the domain of Swiss law. Although effective content-based retrieval of spoken material is still an area of ongoing research effort, we have already established the possibility of cross-language access to speech data with our similarity thesaurus approach.

Although the construction of a similarity thesaurus relies on the availability of a multilingual corpus, we have not yet had problems in locating such a resource in the domains to which we have so far applied this technology. It is also important to note that all approaches to cross-language information retrieval currently under investigation require some form of language resource, either corpus resources or lexical resources such as transfer dictionaries. In fact, as digital libraries grow and as more and more information becomes available in electronic form, the potential problems of finding suitable corpora for building similarity thesauri will become less and less. This is a trend that we are already witnessing.

As digital libraries expand and become accessible to wider audiences without regard to geographic boundaries, it seems that technologies which enable cross-language access, as presented here, will become increasingly important and will serve to at least reduce the barriers of language, so ensuring maximal accessibility to digital repositories for a much wider audience of users.

Acknowledgements

We would like to acknowledge the input of several of our colleagues to this work, especially Martin Braschler, Jean Paul Ballerini, Chantal Sierro and Martin Wechsler. We also thank the Schweizerische Depeschen Agentur (SDA), Herrn Nicola Kessler of the office of the Swiss 'Bundeskanzlei', and the Swiss federal court of justice for making their data available to us for this research.

References

1. L. Ballesteros, W. B. Croft (1996). Dictionary-based Methods for Cross-lingual Information Retrieval, *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*.
2. M. Davis, T. Dunning (1995). A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval. *Proceedings of the Fourth Text Retrieval Conference (TREC4)*.
3. D. Hull, G. Grefenstette (1996). Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland*, pp. 49–57.
4. LDC (1995). *CELEX-2 Lexical Database*, Linguistic Data Consortium, 3615 Market Street, Suite 200, Philadelphia, PA 19104-2608, USA, ldc@ldc.upenn.edu.
5. M. Littman, S. Dumais, T. Landauer (1997). Automatic Cross-linguistic Information Retrieval using Latent Semantic Indexing. *Cross Language Information Retrieval* (To appear), G. Grefenstette, ed.

6. E. Mittendorf, P. Schäuble, P. Sheridan (1995). Applying Probabilistic Term Weighting to OCR Text in the case of a Large Alphabetic Library Catalogue. *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, pp. 328–335.
7. D. Oard (1997). *Speech-Based Information Retrieval for Digital Libraries*, Tech. Rep. CLIS-TR-97-05 LAMP-TR-015, College of Library and Information Services, University of Maryland, College Park, MD 20742.
8. PhonDat (1993). *Acoustic Database of Spoken Standard German (CD-ROM)*, Institut fur Phonetik und Sprachliche Kommunikation, Schellingstr. 3, D–80799 Munich 40, Germany.
9. Y. Qiu (1995). *Automatic Query Expansion Based on a Similarity Thesaurus*, PhD thesis, Swiss Federal Institute of Technology.
10. Y. Qiu, H. Frei (1993). Concept Based Query Expansion. *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, pp. 160–169.
11. J. Rocchio (1971). Relevance Feedback in Information Retrieval. *The SMART Retrieval System—Experiments in Automatic Document Processing*, G. Salton, ed., Prentice-Hall Inc., ch. 14, pp. 313–323.
12. P. Sheridan, J. P. Ballerini (1996). Experiments in Multilingual Information Retrieval using the SPIDER System. *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 58–65.
13. P. Sheridan, J. P. Ballerini, P. Schäuble (1996). Building a Large Multilingual Test Collection From Comparable News Documents. *Cross Language Information Retrieval* (T, appear). G. Grefenstette, ed.
14. P. Sheridan, P. Schäuble (1997). Cross-Language Information Retrieval in a Multilingual Legal Domain. *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, Pisa, Italy.
15. P. Sheridan, M. Wechsler, P. Schäuble (1997). Cross-Language Speech Retrieval: Establishing a Baseline Performance. *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA.
16. M. Wechsler, P. Schäuble (1995). Speech retrieval based on automatic indexing. *Proceedings of the Final Workshop on Multimedia Information Retrieval* (MIRO'95). I. Ruthven, ed., Electronic Workshops in Computing, Glasgow, Springer.
17. ——(1997). Metadata for Content Based Retrieval of Speech Recordings. *Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data*, W. Klas and A. Shet, eds., MacGraw-Hill.
18. S. Young, P. Woodland, and W. Byrne (1993). *HTK Version 1.5: User, Reference & Programmer Manual*, Entropic Cambridge Research Laboratory, Sheraton House, Castle Park, Cambridge CB3 OAX, England.